Loss Data Analytics, Second Edition

An open text authored by the Actuarial Community

Contents

Preface

Date: 19 September 2024

Book Description

Loss Data Analytics is an interactive, online, freely available text.

- The online version contains many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote *deeper learning*.
- A subset of the book is available for *offline reading* in pdf and EPUB formats.
- The online text will be available in multiple languages to promote access to a *worldwide audience*.

What will success look like?

The online text will be freely available to a worldwide audience. The online version will contain many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote deeper learning. Moreover, a subset of the book will be available in pdf format for low-cost printing. The online text will be available in multiple languages to promote access to a worldwide audience.

How will the text be used?

This book will be useful in actuarial curricula worldwide. It will cover the loss data learning objectives of the major actuarial organizations. Thus, it will be suitable for classroom use at universities as well as for use by independent learners seeking to pass professional actuarial examinations. Moreover, the text will also be useful for the continuing professional development of actuaries and other professionals in insurance and related financial risk management industries.

Why is this good for the profession?

An online text is a type of open educational resource (OER). One important benefit of an OER is that it equalizes access to knowledge, thus permitting a broader community to learn about the actuarial profession. Moreover, it

has the capacity to engage viewers through active learning that deepens the learning process, producing analysts more capable of solid actuarial work.

Why is this good for students and teachers and others involved in the learning process? Cost is often cited as an important factor for students and teachers in textbook selection (see a recent post on the $400 textbook). Students will also appreciate the ability to "carry the book around" on their mobile devices.

Why loss data analytics?

The intent is that this type of resource will eventually permeate throughout the actuarial curriculum. Given the dramatic changes in the way that actuaries treat data, loss data seems like a natural place to start. The idea behind the name *loss data analytics* is to integrate classical loss data models from applied probability with modern analytic tools. In particular, we recognize that big data (including social media and usage based insurance) are here to stay and that high speed computation is readily available.

Project Goal

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. To get involved, please visit our Open Actuarial Textbooks Project Site.

Acknowledgements

Edward Frees acknowledges the John and Anne Oros Distinguished Chair for Inspired Learning in Business which provided seed money to support the project. Frees and his Wisconsin colleagues also acknowledge a Society of Actuaries Center of Excellence Grant that provided funding to support work in dependence modeling and health initiatives. Wisconsin also provided an education innovation grant that provided partial support for the many students who have worked on this project.

We acknowledge the Society of Actuaries for permission to use problems from their examinations.

We thank Rob Hyndman, Monash University, for allowing us to use his excellent style files to produce the online version of the book.

We thank Yihui Xie and his colleagues at Rstudio for the R bookdown package that allows us to produce this book.

We also wish to acknowledge the support and sponsorship of the Interna-

tional Association of Black Actuaries in our joint efforts to provide actuarial educational content to all.

IABA
INTERNATIONAL ASSOCIATION OF BLACK ACTUARIES

Contributors

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. The following contributors have taken a leadership role in developing *Loss Data Analytics*.

Zeinab Amin

- **Zeinab Amin** is a Professor at the Department of Mathematics and Actuarial Science and Associate Provost for Assessment and Accreditation at the American University in Cairo (AUC). Amin holds a PhD in Statistics and is an Associate of the Society of Actuaries. Amin is the recipient of the 2016 Excellence in Academic Service Award and the 2009 Excellence in Teaching Award from AUC. Amin has designed and taught a variety of statistics and actuarial science courses. Amin's current area of research includes quantitative risk assessment, reliability assessment, general statistical modelling, and Bayesian statistics.
- **Katrien Antonio**, KU Leuven

Jean-François Bégin

- **Jean-François Bégin** is an Assistant Professor in the Department of Statistics and Actuarial Science at Simon Fraser University in British Columbia, Canada. Bégin holds a PhD in Financial Engineering from HEC Montréal, Canada, and is a Fellow of the Society of Actuaries and of the Canadian Institute of Actuaries. His current research interests include financial modelling, financial econometrics, Bayesian statistics, filtering methods, credit risk, option pricing, and pension economics. Bégin has designed and taught a variety of actuarial finance and actuarial communication courses.
- **Jan Beirlant**, KU Leuven

Arthur Charpentier

- **Arthur Charpentier** is a professor in the Department of Mathematics at the Université du Québec á Montréal. Prior to that, he worked at a large general insurance company in Hong Kong, China, and the French Federation of Insurers in Paris, France. He received a MS on mathematical economics at Université Paris Dauphine and a MS in actuarial science at ENSAE (National School of Statistics) in Paris, and a PhD degree from KU Leuven, Belgium. His research interests include econometrics, applied probability and actuarial science. He has published several books (the most recent one on *Computational Actuarial Science with R*, CRC) and papers on a variety of topics. He is a Fellow of the French Institute of Actuaries, and was in charge of the 'Data Science for Actuaries' program from 2015 to 2018.

Curtis Gary Dean

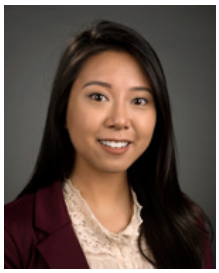
- **Curtis Gary Dean** is the Lincoln Financial Distinguished Professor of Actuarial Science at Ball State University. He is a Fellow of the Casualty Actuarial Society and a CFA charterholder. He has extensive practical experience as an actuary at American States Insurance, SAFECO, and Travelers. He has served the CAS and actuarial profession as chair of the Examination Committee, first editor-in-chief for *Variance: Advancing the Science of Risk*, and as a member of the Board of Directors and the Executive Council. He contributed a chapter to *Predictive Modeling Applications in Actuarial Science* published by Cambridge University Press.

Edward (Jed) Frees

- **Edward (Jed) Frees** is an emeritus professor, formerly the Hickman-Larson Chair of Actuarial Science at the University of Wisconsin-Madison. He is a Fellow of both the Society of Actuaries and the American Statistical Association. He has published extensively (a four-time winner of the Halmstad and Prize for best paper published in the actuarial literature) and has written three books. He also is a co-editor of the two-volume series *Predictive Modeling Applications in Actuarial Science* published by Cambridge University Press.

Guojun Gan

- **Guojun Gan** is an associate professor in the Department of Mathematics at the University of Connecticut, where he has been since August 2014. Prior to that, he worked at a large life insurance company in Toronto, Canada for six years. He received a BS degree from Jilin University, Changchun, China, in 2001 and MS and PhD degrees from York University, Toronto, Canada, in 2003 and 2007, respectively. His research interests include data mining and actuarial science. He has published several books and papers on a variety of topics, including data clustering, variable annuity, mathematical finance, applied statistics, and VBA programming.



Lisa Gao

- **Lisa Gao** is a PhD candidate in the Risk and Insurance department at the University of Wisconsin-Madison. She holds a BMath in Actuarial Science and Statistics from the University of Waterloo and is an Associate of the Society of Actuaries.
- **José Garrido**, Concordia University

Lei (Larry) Hua

- **Lei (Larry) Hua** is an Associate Professor of Actuarial Science at Northern

Illinois University. He earned a PhD degree in Statistics from the University of British Columbia. He is an Associate of the Society of Actuaries. His research work focuses on multivariate dependence modeling for non-Gaussian phenomena and innovative applications for financial and insurance industries.

Noriszura Ismail

- **Noriszura Ismail** is a Professor and Head of Actuarial Science Program, Universiti Kebangsaan Malaysia (UKM). She specializes in Risk Modelling and Applied Statistics. She obtained her BSc and MSc (Actuarial Science) in 1991 and 1993 from University of Iowa, and her PhD (Statistics) in 2007 from UKM. She also passed several papers from Society of Actuaries in 1994. She has received several research grants from Ministry of Higher Education Malaysia (MOHE) and UKM, totaling about MYR1.8 million. She has successfully supervised and co-supervised several PhD students (13 completed and 11 on-going). She currently has about 180 publications, consisting of 88 journals and 95 proceedings.

Joseph H.T. Kim

- **Joseph H.T. Kim**, Ph.D., FSA, CERA, is Associate Professor of Applied Statistics at Yonsei University, Seoul, Korea. He holds a Ph.D. degree in Actuarial Science from the University of Waterloo, at which he taught as Assistant Professor. He also worked in the life insurance industry. He has published papers in *Insurance Mathematics and Economics*, *Journal of Risk and Insurance*, *Journal of Banking and Finance*, *ASTIN Bulletin*, and *North American Actuarial Journal*, among others.

Nii-Armah Okine

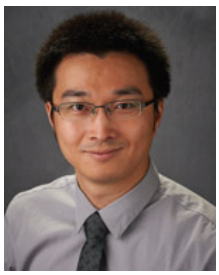
- **Nii-Armah Okine** is an assistant professor at the Mathematical Sciences Department at Appalachian State University. He holds a Ph.D. in Business (Actuarial Science) from the University of Wisconsin - Madison and obtained his master's degree in Actuarial science from Illinois State University. His research interest includes micro-level reserving, joint longitudinal-survival modeling, dependence modeling, micro-insurance, and machine learning.

Rajesh (Raj) Sahasrabuddhe

- **Rajesh (Raj) Sahasrabuddhe** is a Partner and Philadelphia Office Leader with Oliver Wyman Actuarial Consulting. Raj is a Fellow of the Casualty Actuarial Society (CAS), an Associate of the Canadian Institute of Actuaries, and a Member of the American Academy of Actuaries. Raj has been an active volunteer with CAS Admissions committees throughout his career, including a term as Chairperson of the Syllabus Committee from 2010 to 2013. He currently serves on the MAS-II Examination Committee. He has authored or co-authored papers that have appeared on syllabi for both the CAS and Society of Actuaries.

Emine Selin Sarıdaş

- **Emine Selin Sarıdaş** is a doctoral candidate in the Statistics department of Mimar Sinan University. She holds a bachelor degree in Actuarial Science with a minor in Economics and a master degree in Actuarial Science from Hacettepe University. Her research interest includes dependence modeling, regression, loss models and life contingencies.

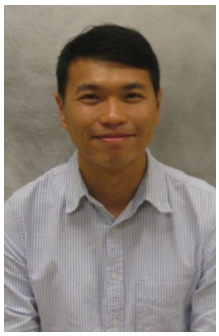


Peng Shi

- **Peng Shi** is an associate professor in the Risk and Insurance Department at the Wisconsin School of Business. He is also the Charles & Laura Albright Professor in Business and Finance. Professor Shi is an Associate of the Casualty Actuarial Society (ACAS) and a Fellow of the Society of Actuaries (FSA). He received a Ph.D. in actuarial science from the University of Wisconsin-Madison. His research interests are problems at the intersection of insurance and statistics. He has won several research awards, including the Charles A. Hachemeister Prize, the Ronald Bornhuetter Loss Reserve Prize, and the American Risk and Insurance Association Prize.

Nariankadu D. Shyamalkumar (Shyamal)

- **Nariankadu D. Shyamalkumar (Shyamal)** is an associate professor in the Department of Statistics and Actuarial Science at The University of Iowa. He is an Associate of the Society of Actuaries, and has volunteered in various elected and non-elected roles within the SoA. Having a broad theoretical interest as well as interest in computing, he has published in prominent actuarial, computer science, probability theory, and statistical journals. Moreover, he has worked in the financial industry, and since then served as an independent consultant to the insurance industry. He has experience educating actuaries in both Mexico and the US, serving in the roles of directing an undergraduate program, and as a graduate adviser for both masters and doctoral students.



Jianxi Su

- **Jianxi Su** is an Assistant Professor at the Department of Statistics at Purdue University. He is the Associate Director of Purdue's Actuarial Science. Prior to joining Purdue in 2016, he completed the PhD at York University (2012-2015). He obtained the Fellow of the Society of Actuaries (FSA) in 2017. His research expertise are in dependence modelling, risk management, and pricing. During the PhD candidature, Jianxi also worked as a research associate at the Model Validation and ORSA Implementation team of Sun Life Financial (Toronto office).

Chong It Tan

- **Chong It Tan** is a senior lecturer at Macquarie University in Australia, where he has served as the undergraduate actuarial program director since 2018. He obtained his PhD in 2015 from Nanyang Technological University in Singapore. He is a fully qualified actuary, holding the credentials from both the US Society of Actuaries and Australian Actuaries Institute. His major research interests are mortality modelling, longevity risk management and bonus-malus systems.

Tim Verdonck

- **Tim Verdonck** is associate professor at the University of Antwerp. He has a degree in Mathematics and a PhD in Science: Mathematics, obtained at the University of Antwerp. During his PhD he successfully took the Master in Insurance and the Master in Financial and Actuarial Engineering, both at KU Leuven. His research focuses on the adaptation and application of robust statistical methods for insurance and finance data.

Krupa Viswanathan

- **Krupa Viswanathan** is an Associate Professor in the Risk, Insurance and Healthcare Management Department in the Fox School of Business, Temple University. She is an Associate of the Society of Actuaries. She teaches courses in Actuarial Science and Risk Management at the undergraduate and graduate levels. Her research interests include corporate governance of insurance companies, capital management, and sentiment analysis. She received her Ph.D. from The Wharton School of the University of Pennsylvania.

Reviewers

Our goal is to have the actuarial community author our textbooks in a collaborative fashion. Part of the writing process involves many reviewers who generously donated their time to help make this book better. They are:

- Yair Babab
- David Back, Liberty Mutual
- Chunsheng Ban, Ohio State University
- Vytaras Brazauskas, University of Wisconsin - Milwaukee
- Yvonne Chueh, Central Washington University
- Chun Yong Chew, Universiti Tunku Abdul Rahman (UTAR)
- Benjamin Côté, Université Laval
- Eren Dodd, University of Southampton
- Gordon Enderle, University of Wisconsin - Madison
- Rob Erhardt, Wake Forest University
- Runhun Feng, University of Illinois
- Brian Hartman, Brigham Young University
- Liang (Jason) Hong, University of Texas at Dallas
- Fei Huang, Australian National University
- Hirokazu (Iwahiro) Iwasawa

- Himchan Jeong, University of Connecticut
- Min Ji, Towson University
- Paul Herbert Johnson, University of Wisconsin - Madison
- Dalia Khalil, Cairo University
- Samuel Kolins, Lebonan Valley College
- Andrew Kwon-Nakamura, Zurich North America
- Ambrose Lo, University of Iowa
- Mélina Mailhot, Concordia University
- Mark Maxwell, University of Texas at Austin
- Tatjana Miljkovic, Miami University
- Bell Ouelega, American University in Cairo
- Zhiyu (Frank) Quan, University of Connecticut
- Jiandong Ren, Western University
- Margie Rosenberg, University of Wisconsin - Madison
- Rajesh V. Sahasrabuddhe, Oliver Wyman
- Sherly Paola Alfonso Sanchez, Universidad Nacional de Colombia
- Ranee Thiagarajah, Illinois State University
- Ping Wang, Saint Johns University
- Chengguo Weng, University of Waterloo
- Toby White, Drake University
- Michelle Xia, Northern Illinois University
- Di (Cindy) Xu, University of Nebraska - Lincoln
- Lina Xu, Columbia University
- Lu Yang, University of Amsterdam
- Chun Yong
- Jorge Yslas, University of Copenhagen
- Jeffrey Zheng, Temple University
- Hongjuan Zhou, Arizona State University

Other Collaborators

- Alyaa Nuval Binti Othman, Aisha Nuval Binti Othman, and Khairina (Rina) Binti Ibraham were three of many students at the Univeristy of Wiscinson-Madison that helped with the text over the years.
- Maggie Lee, Macquarie University, and Anh Vu (then at University of New South Wales) contributed the end of the section quizzes.
- Jeffrey Zheng, Temple University, Lu Yang (University of Amsterdam), and Paul Johnson, University of Wisconsin-Madison, led the work on the glossary.

Version Number

- This is **Version 2.0**, October 2024. Edited by Hélène Cossette, Edward (Jed) Frees, Brian Hartman, and Tim Higgins.
- Version 1.1, August 2020. Edited by Edward (Jed) Frees and Paul Johnson.
- Version 1.0, January 2020, was edited by Edward (Jed) Frees.

You can also access pdf and epub (current and older) versions of the text in our Offline versions of the text.

For our Readers

We hope that you find this book worthwhile and even enjoyable. For your convenience, at our Github Landing site (https://openacttexts.github.io/), you will find links to the book that you can (freely) download for offline reading, including a pdf version (for Adobe Acrobat) and an EPUB version suitable for mobile devices. Data for running our examples are available at the same site.

In developing this book, we are emphasizing the online version that has lots of great features such as a glossary, code and solutions to examples that you can be revealed interactively. For example, you will find that the statistical code is hidden and can only be seen by clicking on terms such as

We hide the code because we don't want to insist that you use the `R` statistical software (although we like it). Still, we encourage you to try some statistical code as you read the book – we have opted to make it easy to learn `R` as you go. We have set up a separate R Code for Loss Data Analytics site to explain more of the details of the code.

Like any book, we have a set of notations and conventions. It will probably save you time if you regularly visit our Appendix Chapter **??** to get used to ours.

Freely available, interactive textbooks represent a new venture in actuarial education and we need your input. Although a lot of effort has gone into the development, we expect hiccoughs. Please let your instructor know about opportunities for improvement, write us through our project site, or contact chapter contributors directly with suggested improvements.

1

Glossary

Term	Definition	Section
analytics	Analytics is the process of using data to make decisions.	1.1
renters insurance	Renters insurance is an insurance policy that covers the contents of an apartment or house that you are renting.	1.1
automobile insurance	An insurance policy that covers damage to your vehicle, damage to other vehicles in the accident, as well as medical expenses of those injured in the accident.	1.1
casualty insurance	Causalty insurance is a form of liability insurance providing coverage for negligent acts and omissions. examples include workers compensation, errors and omissions, fidelity, crime, glass, boiler, and various malpractice coverages.	1.1
commercial insurance		1.1
term	The duration of an insurance contract	1.1
insurance claim	An insurance claim is the compensation provided by the insurer for incurred hurt, loss, or damage that is covered by the policy.	1.1
homeowners insurance	Homeowners insurance is an insurance policy that covers the contents and property of a building that is owned by you or a friend.	1.1
property insurance	Property insurance is a policy that protects the insured against loss or damage to real or personal property. the cause of loss might be fire, lightening, business interruption, loss of rents, glass breakage, tornado, windstorm, hail, water damage, explosion, riot, civil commotion, rain, or damage from aircraft or vehicles.	1.1
non-life	Non-life insurance is any type of insurance where payments are not based on the death (or survivorship) of a named insured. examples include automobile, homeowners, and so on. also known as property and casualty or general insurance.	1.1

(continued)

Term	Definition	Section
life insurance	Life insurance is a contract where the insurer promises to pay upon the death of an insured person. the person being paid is the beneficiary.	1.1
personal insurance	Insurance purchased by a person	1.1
loss adjustment expenses	Loss adjustment expenses are costs to the insurer that are directly attributable to settling a claims. for example, the cost of an adjuster is someone who assess the claim cost or a lawyer who becomes involve in settling an insurer's legal obligation on a claim	1.2
unallocated	Unallocated loss adjustment expenses are costs that can only be indirectly attributed to claim settlement; for example, the cost of an office to support claims staff	1.2
allocated	Allocated loss adjustment expenses, sometimes known by the acronym alea, are costs that can be directly attributed to settling a claim; for example, the cost of an adjuster	1.2
underwriting	Underwriting is the process where the company makes a decision as to whether or not to take on a risk.	1.2
loss reserving	A loss reserve is an estimate of liability indicating the amount the insurer expects to pay for claims that have not yet been realized. this includes losses incurred but not yet reported (ibnr) and those claims that have been reported claims that haven't been paid (known by the acronym rbns for reported but not settled).	1.2
risk classification	Risk classification is the process of grouping policyholders into categories, or classes, where each insured in the class has a risk profile that is similar to others in the class.	1.2
retrospective premiums	The process of determining the cost of an insurance policy based on the actual loss experience determined as an adjustment to the initial premium payment.	1.2
claims adjustment	Claims adjustment is the process of determining coverage, legal liability, and settling claims.	1.2
claims leakage	Claims leakage respresents money lost through claims management inefficiencies.	1.2
adjuster	An adjuster is a person who investigates claims and recommends settlement options based on estimates of damage and insurance policies held.	1.2

(continued)

Term	Definition	Section
dividends	A dividend is the refund of a portion of the premium paid by the insured from insurer surplus.	1.2
indemnification	Indemnification is the compensation provided by the insurer.	1.3
rating variables	Rating variables are the components of an insurance pricing formula. they can include numeric variables (like values, revenue, or area) and classification variables (like location, type of vehicle, or type of occupancy.)	1.3
frequency	Count random variables that represent the number of claims	2.1
severity	The amount, or size, of each payment for an insured event	2.1
probability mass function (pmf)	A function that gives the probability that a discrete random variable is exactly equal to some value	2.1
distribution function	The chance that the random variable is less than or equal to x, as a function of x	2.1
mean	Average	2.1
moments	The rth moment of a list is the average value of the random variable raised to the rth power	2.1
survival function	The probability that the random variable takes on a value greater than a number x	2.1
moment generating function (mgf)	The mgf of random variable n is defined the expectation of exp(tn), as a function of t	2.2
probability generating function (pgf)	For a random variable n, its pgf is defined as the expectation of s^n, as a function of s	2.2
convex hulls	The convex hull of a set of points x is the smallest convex set that contains x	2.2
risk classes	The formation of different premiums for the same coverage based on each homogeneous group's characteristics.	2.2
binomial distribution	A random variable has a binomial distribution (with parameters m and q) if it is the number of "successes" in a fixed number m of independent random trials, all of which have the same probability q of resulting in "success."	2.2
binary outcomes	Outcomes whose unit can take on only two possible states, traditionally labeled as 0 and 1	2.2

(continued)

Term	Definition	Section
m-convolution	The addition of m independent random variables	2.2
poisson distribution	A discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event	2.2
negative binomial distribution	The number of successes until we observe the rth failure in independent repetitions of an experiment with binary outcomes	2.2
overdispersed	The presence of greater variability (statistical dispersion) in a data set than would be expected based on a given statistical model	2.2
underdispersed	There was less variation in the data than predicted	2.2
(a, b, 0) class	The poisson, binomial and negative binomial distributions	2.3
maximum likelihood estimator (mle)	The possible value of the parameter for which the chance of observing the data largest	2.4
local extrema	The largest and smallest value of the function within a given range	2.4
central limit theorem (clt)	In some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.	2.4
newton's method	A root-finding algorithm which produces successively better approximations to the roots of a real-valued function	2.4
robust	Resistant to errors in the results, produced by deviations from assumptions	2.4
explanatory variables	In regression, the explanatory variable is the one that is supposed to "explain" the other.	2.5
regression analysis	A set of statistical processes for estimating the relationships among variables	2.5
homogeneous	Units of exposure that face approximately the same expected frequency and severity of loss.	2.5
(a,b,1)	A count distribution with probabilities satisfying p_k/p_{k-1}=a+b/k, for some some constants a and b and k>=2	2.5

(continued)

Term	Definition	Section
zero truncation	Zero modification of a count distribution such that it assigns zero probability to zero count	2.5
degenerate distribution	A deterministic distribution and takes only a single value	2.5
convex combination	A linear combination of points where all coefficients are non-negative and sum to 1	2.5
convex function	A real-valued function defined on an interval is called convex if the line segment between any two points on the graph of the function lies above or on the graph.	2.6
mixture distribution	The probability distribution of a random variable that is derived from a collection of other random variables as follows: first, a random variable is selected by chance from the collection according to given probabilities of selection, and then the value of the selected random variable is realized	2.6
chi-square distribution	The chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables	2.7
aic	A goodness of fit measure of a statistical model that describes how well it fits a set of observations.	2.7
pearson's chi-square test	A statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance	2.7
multinomial likelihood	The multinomial distribution models the probability of counts for rolling a k-sided die n times	2.7
aggregate losses	Aggregate claims, or total claims observed in the time period	3
liability insurance	Insurance that compensates an insured for loss due to legal liability towards others	3
mixture distribution	A weighted average of other distributions, which may be continuous or discrete	3
continuous random variable	Random variable which can take infinitely many values in its specified domain	3.1
raw moment	The kth moment of a random variable x is the average (expected) value of x^k	3.1
central moment	The kth central moment of a random variable x is the expected value of (x-its mean)^k	3.1
skewness	Measure of the symmetry of a distribution, 3rd central moment/standard deviation^3	3.1

(continued)

Term	Definition	Section
kurtosis	Measure of the peaked-ness of a distribution, 4th central moment/standard deviation^4	3.1
expected value	Average	3.1
exponential distribution	A single parameter continous probability distribution that is defined by its rate parameter	3.1
independent	Two variables are independent if conditional information given about one variable provides no information regarding the other variable	3.1
percentile	The pth percentile of a random variable x is the smallest value x_p such that the probability of not exceeding it is p%	3.1
chi-square distribution	A common distribution used in chi-square tests for determining goodness of fit of observed data to a theorized distribution	3.2
light tailed distribution	A distribution with thinner tails than the benchmark exponential distribution	3.2
pareto distribution	A heavy-tailed and positively skewed distribution with 2 parameters	3.2
hazard function	Ratio of the probability density function and the survival function: f(x)/s(x), and represents an instantaneous probability within a small time frame	3.2
weibull distribution	A positively skewed continuous distribution with 2 parameters that can have an increasing or decreasing hazard function depending on the shape parameter	3.2
generalized beta distribution of the second kind	A 4-parameter flexible distribution that encompasses many common distributions	3.2
parametric distributions	Probability distribution defined by a fixed set of parameters	3.3
transformation	A function or method that turns one distribution into another	3.3
distribution function technique	A transformation technique that involves finding the cdf of the transformed distribution through its relation with the original cdf	3.3
change-of-variable technique	A transformation technique that involves finding the pdf of the transformed distribution through its relation with the original pdf using inverse functions	3.3

(continued)

Term	Definition	Section
moment-generating function technique	A transformation technique that uses moment generating functions properties to determine the mgf of a linear combination of variables	3.3
lognormal distribution	A heavy-tailed, positively skewed 2-parameter continuous distribution such that the natural log of the random variable is normally distributed with the same parameter values	3.3
reliability data	A dataset consisting of failure times for failed units and run times for units still functioning	3.3
power transformation	A transformation type that involves raising a random variable to a power	3.3
exponential transformation	A transformation type that involves raising a random variable in the exponent	3.3
mixing parameters	Proportion weight given to each subpopulation in a mixture	3.3
heterogeneous population	A dataset where the subpopulations are represented by separate distinct distributions	3.3
finite mixture	A mixture distribution with a finite k number of subpopulations	3.3
continuous mixture	A mixture distribution with an infinite number of subpopulations, where the mixing parameter is itself a continuous distribution	3.3
conditional distribution	A probability distribution that applies to a subpopulation satisfying the condition	3.3
unconditional distribution	A probability distribution independent of any another imposed conditions	3.3
prior distribution	A probability distribution assigned prior to observing additional data	3.3
scale distribution	A distribution with the property that multiplying all values by a constant leads to the same distribution family with only the scale parameter changed	3.3
moral hazard	Situation where an insured is more likely to be risk seeking if they do not bear sufficient consequences for a loss	3.4
payment per loss	Amount insurer pays when a loss occurs and can be 0	3.4
payment per payment	Amount insurer pays given a payment is needed and is greater than 0	3.4

(continued)

Term	Definition	Section
left censored	Values below a threshold d are not ignored but converted to 0	3.4
left truncated	Values below a threshold d are not reported and unknown	3.4
loss elimination ratio (ler)	% decrease of the expected payment by the insurer as a result of the deductible	3.4
franchise deductible	Insurer pays nothing for losses below the deductible, but pays the full amount for any loss above the deductible	3.4
limit of coverage	Policy limit, or maximum contractual financial obligation of the insurer for a loss	3.4
group insurance	Insurance provided to groups of people to take advantage of lower administrative costs vs. individual policies	3.4
growth factor	Multiplicative factor applied to a distribution to account for the impact of inflation, typically (1+rate)	3.4
cedent	Party that is transferring the risk to a reinsurer	3.4
excess of loss coverage	Contract where an insurer pays all claims up to a specified amount and then the reinsurer pays claims in excess of stated reinsurance deductible	3.4
retention	Maximum amount payable by the primary insurer in a reinsurance arrangement	3.4
right censored variable	Values above a threshold u are not ignored but converted to u	3.4
reinsurance	A transaction where the primary insurer buys insurance from a re-insurer who will cover part of the losses and/or loss adjustment expenses of the primary insurer	3.4
method of maximum likelihood	Statistical method used to derive the parameter values from data that maximize the probability of observing the data given the parameters	3.5
grouped data	Data bucketed into categories with ranges, such as for use in histograms or frequency tables	3.5
large-sample properties	Asymptotic properties of a distribution as the amount of data increases towards infinity	3.5
asymptotic variance	Variability of the distribution of an estimator as the amount of data increases towards infinity	3.5
delta method	Statistical method used to approximate the asymptotic variance for a function based on parameters whose asymptotic variance can be determined	3.5

(continued)

Term	Definition	Section
log-likelihood function	Natural log of the likelihood function	3.5
covariance matrix	Matrix where the (i,j)^th element represents the covariance between the ith and jth random variables	3.5
complete data	Data where each individual observation is known, and no values are censored, truncated, or grouped	3.5
parametric	Distributional assumptions made on the population from which the data is drawn, with properties defined using parameters.	4.1
nonparametric	No distributional assumptions are made on the population from which the data is drawn.	4.1
sampling scheme	How the data is obtained from the population and what data is observed.	4.1
unbiased	An estimator that has no bias, that is, the expected value of an estimator equals the parameter being estimated.	4.1
plug-in principle	The plug-in principle or analog principle of estimation proposes that population parameters be estimated by sample statistics which have the same property in the sample as the parameters do in the population.	4.1
indicator	A categorical variable that has only two groups. the numerical values are usually taken to be one to indicate the presence of an attribute, and zero otherwise. another name for a binary variable.	4.1
empirical distribution function	The empirical distribution is a non-parametric estimate of the underlying distribution of a random variable. it directly uses the data observations to construct the distribution, with each observed data point in a size-n sample having probability 1/n.	4.1
first quartile	The 25th percentile; the number such that approximately 25% of the data is below it.	4.1
third quartile	The 75th percentile; the number such that approximately 75% of the data is below it.	4.1
quantile	The q-th quantile is the point(s) at which the distribution function is equal to q, i.e. the inverse of the cumulative distribution function.	4.1
smoothed empirical quantile	A quantile obtained by linear interpolation between two empirical quantiles, i.e. data points.	4.1

(continued)

Term	Definition	Section
bandwidth	A small positive constant that defines the width of the steps and the degree of smoothing.	4.1
kernel density estimator	A nonparametric estimator of the density function of a random variable.	4.1
bias-variance tradeoff	The tradeoff between model simplicity (underfitting; high bias) and flexibility (overfitting; high variance).	4.1
model diagnostics	Procedures to assess the validity of a model	4.1
probability-probability (pp) plot	A plot that compares two models through their cumulative probabilities.	4.1
quantile-quantile (qq) plot	A plot that compares two models through their quantiles.	4.1
goodness of fit statistics	A measure used to assess how well a statistical model fits the data, usually by summarizing the discrepancy between the observations and the expected values under the model.	4.1
goodness of fit	A measure used to assess how well a statistical model fits the data, usually by summarizing the discrepancy between the observations and the expected values under the model.	4.1
method of moments	The estimation of population parameters by approximating parametric moments using empirical sample moments.	4.1
percentile matching	The estimation of population parameters by approximating parametric percentiles using empirical quantiles.	4.1
percentile	A 100p-th percentile is the number such that 100 times p percent of the data is below it.	4.1
gini index	A measure for assessing income inequality. it measures the discrepancy between the income and population distributions and is calculated from the lorenz curve.	4.2
model selection	The process of selecting a statistical model from a set of candidate models using data.	4.2
in-sample	A dataset used for analysis and model development. also known as a training dataset.	4.2
out-of-sample	A dataset used for model validation. also known as a test dataset.	4.2

(continued)

Term	Definition	Section
cross-validation	A model validation procedure in which the data sample is partitioned into subsamples, where splits are formed by separately taking each subsample as the out-of-sample dataset.	4.2
model validation	The process of confirming that the proposed model is appropriate.	4.2
data-snooping	Repeatedly fitting models to a data set without a prior hypothesis of interest.	4.2
predictive inference	Preditive inference is the process of using past data observations to predict future observations.	4.2
likelihood function	A function of the likeliness of the parameters in a model, given the observed data.	4.3
ogive estimator	A nonparametric estimator for the distribution function in the presence of grouped data.	4.3
product-limit estimator	A nonparametric estimator of the survival function in the presence of incomplete data. also known as the kaplan-meier estimator.	4.3
risk set	The number of observations that are active (not censored) at a specific point.	4.3
nelson-aalen	A nonparametric estimator of the cumulative hazard function in the presence of incomplete data.	4.3
credibility	An actuarial method of balancing an individual's loss experience and the experience in the overall portfolio to improve ratemaking estimates.	4.4
bayesian	A type of statistical inference in which the model parameters and the data are random variables.	4.4
predictive distribution	The distribution of new data, conditional on a base set of data, under the bayesian framework.	4.4
least squares	A technique for estimating parameters in linear regression. it is a standard approach in regression analysis to the approximate solution of overdetermined systems. in this technique, one determines the parameters that minimize the sum of squared differences between each observation and the corresponding linear combination of explanatory variables.	4.4
markov chain monte carlo (mcmc) simulation	The class of numerical methods that use markov chains to generate draws from a posterior distribution.	4.4

(continued)

Term	Definition	Section
improper prior	A prior distribution in which the sum or integral of the distribution is not finite.	4.4
confidence interval	Another term for interval estimate. unlike a point estimate, it gives a range of reliability for approximating a parameter of interest.	4.4
decision analysis	Bayesian decision theory is the study of an agent's choices, which is informed by bayesian probability.	4.4
conjugate distributions	Distributions such that the posterior and the prior come from the same family of distributions.	4.4
credibility interval	A summary of the posterior distribution of parameters under the bayesian framework.	4.4
prior distribution	The distribution of the parameters prior to observing data under the bayesian framework.	4.4
exposure	A measure of the rating units for which rates are applied to determine the premium. for example, exposures may be measured on a per unit basis (e.g. a family with auto insurance under one contract may have an exposure of 2 cars) or per $1,000 of value (e.g. homeowners insurance).	5.1
inflation	Inflation is a sustained increase in the general price level of goods and services over a period of time.	5.1
business line		5.1
individual risk model	A modeling approach for aggregate losses in which the loss from each individual contract is considered.	5.1
collective risk model	A modeling approach for aggregate losses in which the aggregate loss is represented in terms of a frequency distribution and a severity distribution.	5.1
coverage	Insurance coverage is the amount of risk or liability that is covered for an individual or entity by an insurance policy.	5.1
frequency distribution	The random number of claims that occur under the collective risk model.	5.1
severity distribution	The randomly distributed amount of each loss under the collective risk model.	5.1
central limit theorem	Given certain conditions, the arithmetic mean of a large number of replications of independent random variables, each with a finite mean and variance, will be approximately normally distributed, regardless of the underlying distribution.	5.2

(continued)

Term	Definition	Section
term life insurance	A term life insurance policy is payable only if death of the insured occurs within a specified time, such as 5 or 10 years, or before a specified age.	5.2
pure endowment	A pure endowment is an insurance policy that is payable at the end of the policy period if the insured is still alive. if the insured has died, there is nothing paid in the form of benefits.	5.2
support	The set of all outcomes for a random variable following some distribution. for example, exponentially distributed random variable x has support x>0.	5.2
convolution	The convolution of probability distributions is the distribution corresponding to the addition of independent random variables.	5.2
law of iterated expectations	A decomposition of the expected value of a random variable into conditional components. specifically, for random variables x and y, e(x) = e[e(x|y)].	5.3
compound distribution	A random variable follows a compound distribution if it is parameterized and contains at least one parameter that is itself a random variable. for example, the tweedie distribution is a compound distribution.	5.3
tweedie distribution	A compound distribution that is a poisson sum of gamma random variables. because it can accommodate a discrete probability mass at zero and a continuous positive component, it is suitable for modeling aggregate insurance claims.	5.3
shape parameter	A numerical parameter of a parametric distribution affecting the shape of a distribution rather than simply shifting it (as a location parameter does) or stretching/shrinking it (as a scale parameter does).	5.3
scale parameter	A numerical parameter of a parametric distribution that stretches/shrinks the distribution without changing its location or shape. the larger the scale parameter, the more spread out the distribution. the scale parameter is also the reciprocal of the rate parameter. for example, the normal distribution has scale parameter \sigma.	5.3
exponential dispersion	A set of distributions that represents a generalisation of the natural exponential family and also plays an important role in generalized linear models.	5.3

(continued)

Term	Definition	Section
generalized linear models	Commonly known by the acronym glm. an extension of the linear regression model where the dependent variable is a member of the linear exponential family. glm encompasses linear, binary, count, and long-tailed, regressions all as special cases.	5.3
exponential family	A family of parametric distributions that are practical for modeling the underlying response variable in generalized linear models. this family includes the normal, bernoulli, poisson, and tweedie distributions as special cases, among many others.	5.3
monte carlo simulation	A computerized statistical model that simulates the effects of various types of uncertainty.	5.4
empirical distribution	The empirical distribution is a non-parametric estimate of the underlying distribution of a random variable. it directly uses the data observations to construct the distribution, with each observed data point in a size-n sample having probability 1/n.	5.4
converge	A type of stochastic convergence for a sequence of random variables x_1,. . . , x_n that approaches some other distribution as n approaches \infty.	5.4
policy limits	A policy limit is the maximum value covered by a policy.	5.5
ground-up loss	The total amount of loss sustained before policy adjustments are made (i.e. before deductions are applied for coinsurance, deductibles, and/or policy limits.)	5.5
per-loss basis	Due to policy modifications (e.g. deductibles), not all losses that occur result in payment. the per-loss basis considers every loss that occurs.	5.5
per-payment basis	Due to policy modifications (e.g. deductibles), not all losses that occur result in payment. the per-payment basis which considers only the losses that result in some payment to the insured.	5.5
memoryless	The memoryless property means that a given probability distribution is independent of its history and what has already elapsed. specifically, random variable x is memoryless if pr(x > s+t | x >= s) = pr(x > t). note that it does not mean x > s+t and x >= s are independent events.	5.5

(continued)

Term	Definition	Section
central limit theorem	The sample mean and sample sum of a random sample of n from a population will converge to a normal curve as the sample size n grows	6.1
simulations	A computer generation of various hypothetical conditions and outputs, based on the model structure provided	6.1
linear congruential generator	Algorithm that yields pseudo-randomized numbers calculated using a linear recursive relationship and a starting seed value	6.1
pseudo-random numbers	Values that appear random but can be replicated by formula	6.1
inverse transform method	Samples a uniform number between 0 and 1 to represent the randomly selected percentile, then uses the inverse of the cumulative density function of the desired distribution to simulate from in order to find the simulated value from the desired distribution	6.1
quantile function	Inverse function for the cumulative density function which takes a percentile value in [0,1] as the input, and outputs the corresponding value in the distribution	6.1
greatest lower bound	Largest value that is less than or equal to a specified subset of values/elements	6.1
universal life insurance	Type of cash value life insurance where the policy's cash value is the excess of premium payments over the cost of insurance, accumulated with interest, with adjustable premiums and coverage over time	6.1
variable life insurance	Type of life insurance whose face value and coverage term can vary depending upon the performance of underlying invested securities	6.1
sampling variability	How much an estimate can vary between samples	6.1
cauchy distribution	A continuous distribution that represents the distribution of the ratio of two independent normally random variables, where the denominator distribution has mean zero	6.1
kolmogorov-smirnov test	A nonparametric statistical test used to determine if a data sample could come from a hypothesized continuous probability distribution	6.1

(continued)

Term	Definition	Section
bootstrap	A method of sampling with replacement from the original dataset to create additional simulated datasets of the same size as the original	6.2
nonparametric approach	A statistical method where no assumption is made about the distribution of the population	6.2
parametric approach	A statistical method where a prior assumption is made about the distribution or model form	6.2
bias	The difference between the expected value of an estimator and the parameter being estimated. bias is an estimation error that does not become smaller as one observes larger sample sizes.	6.2
bias-corrected estimator	If an estimator is known to be consistently biased in a manner, it can be corrected using a factor to be come less biased or unbiased	6.2
jensen inequality	For a convex function f(x), f(expected value of x) <= expected value of f(x)	6.2
natural estimator	An estimator that uses the sample moments as the estimators for the population	6.2
percentile bootstrap interval	Confidence interval for the parameter estimates determined using the actual percentile results from the bootstrap sampling approach, as every bootstrap sample has an associated parameter estimate(s) that can be ranked against the others	6.2
k-fold cross-validation	A type of validation method where the data is randomly split into k groups, and each of the k groups is held out as a test dataset in turn, while the other k-1 gropus are used for distribution or model fitting, with the process repeated k times in total	6.3
leave-one-out cross validation	A special case of k-fold cross validation, where each single data point gets a turn in being the lone hold-out test data point, and n separate models in total are built and tested	6.3
jackknife statistics	To calculate an estimator, leave out each observation in turn, calculate the sample estimator statistic each time, and average over the n separate estimates	6.3
accept-reject mechanism	A sampling method that is used where the random sample is discarded if not within a certain pre-specified range [a, b] and is commonly used when the traditional inverse transform method cannot be easily used	6.4

(continued)

Term	Definition	Section
importance sampling mechanism	Type of sampling method where values in the region of interest can be over-sampled or values outside the region of interest can be under-sampled	6.4
ergodic theorem	Ergodic theory studies the behavior of a dynamical system when it is allowed to run for an extended time	6.5
markov process	A stochastic (time dependent) process that satisfies memorylessness, meaning future predictions of the process can be made solely based on its present state and not the historical path	6.5
invariant measure	Any mathematical measure that is preserved by a function (the mean is an example)	6.5
composants	Component (smaller, self-contained part of larger entity)	6.5
hastings metropolis	A markov chain monte carlo (mcmc) method for random sampling from a probability distribution where values are iteratively generated, with the distribution of the next sample dependent only on the current sample value, and at each iteration, the candidate sample can be either accepted or rejected	6.5
premium	Amount of money an insurer charges to provide the coverage described in the policy	7.1
ratemaking	Process used by insurers to calculate insurance rates, which drive insurance premiums	7.1
insurance rates	Amount of money needed to cover losses, expenses, and profit per one unit of exposure	7.1
insured contingent event	A condition that results in an insurance claim	7.1
expected costs	The cost to an insurer of payments to the insured and allocated loss adjustment expenses (alaes). overhead and profit are not included	7.1
underwriting profit	Profit an insurer derives from providing coverage, excluding investment income	7.1
experience rating	A type of rating plan that uses the insured's historical loss experience as part of the premium determination	7.1
price	A quantity, usually of money, that is exchanged for a good or service	7.1
rates	A rate is the price, or premium, charged per unit of exposure. a rate is a premium expressed in standardized units.	7.1
technical prices		7.1

(continued)

Term	Definition	Section
loss cost	The sum of losses divided by an exposure; it is also known as the pure premium.	7.2
profit loading	A factor or percentage applied to the premium calculation to account for insurer profit in a policy	7.2
indicated change factor	A factor calculated from the loss ratio method that calculates how the rates should change, with factors > 1 indicating an increase and vice versa	7.2
indicated rate	In a rate filing, the amount that the loss experience suggests that the insurer should charge to cover costs.	7.2
credibility	Weight assigned to observed data vs. that assigned to an external or broader-based set of data	7.4
parametric distribution	Model assumption that the sample data comes from a population that can be modeled by a probability distribution with a fixed set of parameters	7.4
commercial business property	Line of business that insures against damage to their buildings and contents due to a covered cause of loss	7.4
continuous variables	Type of variable that can take on any real value	7.4
discrimination	Process of determining premiums on the basis of likelihood of loss. insurance laws prohibit "unfair discrimination".	7.4
rating factor	A rating factor, or rating variable, is a characteristic of the policyholder or risk being insured by which rates vary.	7.4
rating variable	A rating factor, or rating variable, is a characteristic of the policyholder or risk being insured by which rates vary.	7.4
factor	A variable that varies by groups or categories.	7.4
relativity	The difference of the expected risk between a specific level of a rating factor and an accepted baseline value. this difference may be arithmetic or proportional.	7.4
scale distribution	Suppose that y = c x, where x comes from a parametric distribution family and c is a positive constant. the distribution is said to be a scale distribution if (i) the distributions of y and x come from the same family and (ii) only a single parameter differs and that by a factor of c.	7.4

(continued)

Term	Definition	Section
written exposures	Exposure is based off policies written/issued	7.5
earned exposures	Exposure is based off amount exposed to loss for which coverage has been provided	7.5
unearned exposures	Exposure amount for which coverage has not yet been provided	7.5
in force exposures	Exposure amount subject to loss at a particular point in time	7.5
calendar year method	Experience for rating is aggregated based on calendar year, as opposed to other methods such as when a policy term began	7.5
accident date	Date of loss occurrence that gives rise to a claim	7.5
report date	Date when insurer is notified of the claim	7.5
open claim	A claim that has been reported but not yet closed	7.5
mix of business	Different types of policies in an insurer's portfolio	7.5
on-level earned premium	Earned premium of historical policies using the current rate structure	7.5
experience loss ratio	Ratio of experience loss to on-level earned premium in the experience period	7.5
claim	The amount paid to an individual or corporation for the recovery, under a policy of insurance, for loss that comes within that policy.	7.5
incurred but not reported	A claim is said to be incurred but not reported if the insured event occurs prior to a valuation date (and hence the insurer is liable for payment) but the event has not been reported to the insurer.	7.5
closed	A claim is said to be closed when the company deems its financial obligations on the claim to be resolved.	7.5
valuation date	A valuation date is the date at which a company summarizes its financial position, typically quarterly or annually.	7.5
policy year	This is the period between a policy's anniversary dates.	7.5
gini index	The gini index is twice the area between a lorenz curve and a 45 degree line.	7.6
line of equality	45 degree line equating x and y, that represents a perfect alignment in the sample and population distribution	7.6
pp plot	Statistical plot used to assess how close a data sample matches a theorized distribution	7.6

(continued)

Term	Definition	Section
performance curve	A concentration curve is a graph of the distribution of two variables, where both variables are ordered by only one of variables. for insurance applications, it is a graph of distribution of losses versus premiums, where both losses and premiums are ordered by premiums.	7.6
community rating	This generally refers to the premium principle where all risks pay the same amount.	7.6
market conduct regulation	Regulation that ensures consumers obtain fair and reasonable insurance prices and coverage	7.7
government prescribed	Government sets the entire rating system including coverages	7.7
prior approval	Regulator must approve rates, forms, rules filed by insurers before use	7.7
no file	Insurers may use new rates, forms, rules without approval from regulators	7.7
file only	Insurers must file rates, forms, rules for record keeping and use immediately	7.7
rating factors	Characteristics of a risk that help price the insurance contract	8
multiplicative tariff model	A rating method where each rating factor is the product of parameters associated with that rating factor	8
risk characteristics	The distinguishing features of a policy that help determine the expected loss on the policy	8.1
gross insurance premium	Sum of expected losses and expenses and profit on a policy	8.1
adverse selection	A pricing structure that entices riskier individuals to purchase and discourages low-risk individuals from purchasing	8.1
adverse selection spiral	Phenomenon where a book of business deteriorates as it attracts ever-riskier individuals when forced to increase premiums due to losses	8.1
a priori variables	Variables which the insurer has prior knowledge of before the policy inception	8.1
closed-form expressions	A mathematical expression that can be well defined with a formula that has a finite number of operations	8.2
levels	Different outcomes of a categorical variable	8.2
nominal	A categorical variable where the categories do not have a natural order and any numbering is arbitrary	8.2

(continued)

Term	Definition	Section
dummy variables	A variable that takes on a value of 0 or 1 to indicate the absence or presence of a categorical characteristic	8.2
log linear form	Linear regression model where the response variable is the natural log of the expected response value	8.2
base case	The categorical level chosen as the default with all dummy variable indicators of 0	8.2
workers compensation	A no-fault insurance system prescribed by state law where benefits are provided by an employer to an employee due to a job-related injury, including death, resulting from an accident or occupational disease	8.2
exposure bases	The unit of measurement chosen to represent the exposure for a particular risk	8.2
offset	Natural log of the exposure amount that is added to a regression model to account for varying exposures	8.2
tariff	A table or list that contains the rating factors and associated premiums and other risk information	8.3
in-force times	The timeframe during which a policy is active and the insurer is bound by the contractual obligation	8.3
rate parameter	Parameter in certain distributions, such as the exponential, that indicate how quickly the function decays, and it is the reciprocal of the scale parameter	8.3
functional forms	The algebraic relationship between a dependent variable and explanatory variables	8.3
multiplicative form	Relationship where the dependent variable is a product of the explanatory variables	8.3
base tariff cell	The chosen set of rating categories where the rate equals the intercept of the model (the base value)	8.3
relativities	A numerical estimate of value in one category relative to the value in a base classification, typically expressed as a factor	8.3
non-automobile vehicles	Motorized vehicles which are not autos, such as atvs, off-road vehicles, go-carts, etc.	8.3
distributional structure	The manner in which a statistical distribution is parameterized	8.3
information matrix	Matrix that measures the amount of information that an observable random variable x carries about an unknown parameter of a distribution, and is used to calculate covariance matrices of maximum likelihood estimators	8.5

(continued)

Term	Definition	Section
classification rating plan	A rating plan that uses an insured's risk characteristics to determine premium	9.1
credibility weight	The weight assigned to an insured's historical loss experience for the purposes of determining their premium in an experience rating plan	9.1
complement of credibility	The remainder of the weight not assigned to an insured's historical loss experience in the experience rating plan	9.1
class rate	Average rate per exposure for an insured in a particular classification group	9.1
full credibility standard	The threshold of experience necessary to assign 100% credibility to the insured's own experience	9.2
limited fluctuation credibility	A credibility method that attempts to limit fluctuations in its estimates	9.2
cumulative distribution function of the standard normal	Cumulative density function for the normal distribution with mean 0 and standard deviation 1	9.2
buhlmann credibility	A credibility method that uses the amount of experience, expected value of the process variance, and variance of the hypothetical means to determine the credibility weight	9.3
collective mean	The mean estimate of a risk when no loss information about the risk is known	9.3
law of total expectation	The expected value of the conditional expected value of x given y is the same as the expected value of x	9.3
risk parameter	Parameter in a distribution whose value reflects the risk categorization	9.3
expected value of the process variance	Average of the natural variability of observations from within each risk	9.3
variance of the hypothetical means	Variance of the means across different classes, used to determine how similar or different the classes are from one another	9.3
buhlmann-straub credibility	An extension of the buhlmann credibility model that allows for varying exposure by year	9.4

(continued)

Term	Definition	Section
bayes theorem	A probability law that expresses conditional probability of the event a given the event b in terms of the conditional probability of the event b given the event a and the unconditional probability of a	9.5
bayesian inference	A branch of statistics that leverages bayes theorem to update the distribution as more experience becomes available	9.5
gamma-poisson model	A statistical model that assumes the frequency of claims is poisson whose mean has a prior distribution that is a gamma distribution	9.5
exact credibility	A situation where the bayesian credibility estimate matches that of the buhlmann credibility estimate	9.5
beta-binomial model	A statistical model for modeling the probability of an event using the binomial distribution with a probability that has a prior distribution from a beta distribution	9.5
nonparametric estimation	Statistical method that allows the functional form of a fit from data to have no assumed prior distribution, constraints, or parameters	9.5
empirical bayes methods	Credibility methods that estimate the credibility weight without using any assumptions about prior distributions or likelihoods, instead relying only on empirical data	9.5
semiparametric estimation	Credibility method that assumes a distribution for the loss per exposure random variable and otherwise uses empirical data	9.5
portfolios	A collection of contracts	10.1
insurance portfolios	A collection, or aggregation, of insurance contracts	10.1
reinsurers	A company that sells reinsurance	10.1
heavy tailed	A rv is said to be heavy tailed if high probabilities are assigned to large values	10.2
survival function	One minus the distribution function. it gives the probability that a rv exceeds a specific value.	10.2
coherent risk measure	A risk measure that is is subadditive, monontonic, has positive homogeneity, and is translation invariant.	10.3
mean excess loss function	The expected value of a loss in excess of a quantity, given that the loss exceeds the quantity	10.3
risk measure	A measure that summarizes the riskiness, or uncertainty, of a distribution	10.3
value-at-risk	A risk measure based on a quantile function	10.3

(continued)

Term	Definition	Section
ceding company	A company that purchases reinsurance (also known as the reinsured)	10.4
excess of loss	Under an excess of loss arrangement, the insurer sets a retention level for each claim and pays claim amounts less than the level with the reinsurer paying the excess.	10.4
primary insurance	Insurance purchased by a non-insurer	10.4
proportional reinsurance	An agreement between a reinsurer and a ceding company (also known as the reinsured) in which the reinsurer assumes a given percent of losses and premium	10.4
quota share	A proportional treaty where the reinsurer receives a flat percent of the premium for the book of business reinsured and pays a percentage of losses, including allocated loss adjustment expenses. the reinsurer may also pays the ceding company a ceding commission which is designed to reflect the differences in underwriting expenses incurred.	10.4
reinsured	A company that purchases reinsurance (also known as the ceding company)	10.4
retained line	The amount of exposure that the the reinsured retains on a given line in a surplus share reinsurance agreement.	10.4
retention function	A function that maps the insurer portfolio loss into the amount of loss retained by the insurer.	10.4
stop-loss	Under a stop-loss arrangement, the insurer sets a retention level and pays in full total claims less than the level with the reinsurer paying the excess.	10.4
surplus share	A proportional reinsurance treaty that is common in commercial property insurance. a surplus share treaty allows the reinsured to limit its exposure on any one risk to a given amount (the retained line). the reinsurer assumes a part of the risk in proportion to the amount that the insured value exceeds the retained line, up to a given limit (expressed as a multiple of the retained line, or number of lines).	10.4
treaty	A reinsurance contract that applies to a designated book of business or exposures.	10.4
bonus-malus system	A type of rating mechanism where insured premiums are adjusted based on their individual loss experience history	12.1

(continued)

Term	Definition	Section
no claim discount (ncd) system	A type of experience rating where insureds obtain discounts on future years' premiums based on claims-free experience	12.1
hunger for bonus	Phenomenon where insureds under an experience rating system are dissuaded from filing minor claims in order to keep their no-claims discount	12.1
takaful	Co-operative system of reimbursement or repayment in case of loss as an insurance alternative	12.2
markov chain	A stochastic model (time dependent) where the probability of each event depends only on the current state and not the historical path	12.3
transition matrix	Matrix that represents all probabilities for transition from one state to another (could be same state) for a markov chain	12.3
stationary distribution	Probability distribution remains unchanged in the markov chain as time progresses	12.4
ergodic	Irreducible markov chain where it is eventually possible to move from any state to any other state, with positive probability	12.4
irreversible	A markov chain where there does not exist a probability distribution that allows for the chain to be walked backwards in time	12.4
eigenvector	A non-zero vector that changes by only a scalar factor when that linear transformation is applied	12.4
n-step transition probability	Probability of ending in a state j after n periods, starting in state i, where i and j can be the same state	12.4
convergence rate	After n transitions, the sum of variation between the probability in each state vs. the stationary probability	12.4
poisson regression model	Type of regression model used for fitting data with an integral (count) response variable with mean equal to the variance	12.5
negative binomial regression model	Type of regression model used for fitting data with an integral (count) response variable and can account for variance greater than the mean	12.5
overdispersion	Phenomenon where the variance of data is larger than what is modeled	12.5
cross-classified rating classes	Table that combines the effects of multiple rating classifications	12.5

(continued)

Term	Definition	Section
structured data	Data that can be organized into a repository format, typically a database	13.1
unstructured data	Data that is not in a predefined format, most notably text, audio visual	13.1
qualitative data	Data which is non numerical in nature	13.1
quantitative data	Data which is numerical in nature	13.1
ordinal data	Data field with a natural ordering	13.1
interval data	Continuous data which is broken into interval bands with a natural ordering	13.1
key-value databases	Data storage method that stores amd finds records using a unique key hash	13.1
column-oriented databases	Data storage method that stores records by column instead of by row	13.1
document databases	Data storage method that uses the document metadata for search and retrieval, also known as semi-structured data	13.1
data decay	Corruption of data due to hardware failure in the storage device	13.1
reverification	Manual process of checking the integrity of data	13.1
data element analysis	Analysis of the format and definition of each field	13.1
structural analysis	Statistical analysis of the structured data present to detect irregularities	13.1
robust	Statistics which are more unaffected by outliers or small departures from model assumptions	13.2
exploratory data analysis	Approach to analyzing data sets to summarize their main characteristics, using visual methods, descriptive statistics, clustering, dimension reduction	13.2
confirmatory data analysis	Process used to challenge assumptions about the data through hypothesis tests, significance testing, model estimation, prediction, confidence intervals, and inference	13.2
supervised learning methods	Model that predicts a response target variable using explanatory predictors as input	13.2
unsupervised learning methods	Models that work with explanatory variables only to describe patterns or groupings	13.2

(continued)

Term	Definition	Section
classification methods	Supervised learning method where the response is a categorical variable	13.2
regression methods	Classical supervised learning method where the response may be continuous, binary, or a mixture of discrete and continuous	13.2
model flexibility	A measure of model complexity, typically based on the number of estimated parameters	13.2
explanatory modeling	Process where the modeling goal is to identify variables with meaningful and statistically significant relationships and test hypotheses	13.2
predictive modeling	Process where the modeling goal is to predict new observations	13.2
data modeling	Assumes data generated comes from a stochastic data model	13.2
algorithmic modeling	Assumes data generated comes from unknown algorithmic models	13.2
predictive accuracy	Quantitative measure of how well the explanatory variables predict the response outcome	13.2
scripts	A program or sequence of instructions that is executed by another program	13.2
reproducible analysis	Modeling practice where data, code, analyses are published together in a manner so that others may verify the findings	13.2
literate programming	Coding practice where documentation and code are written together	13.2
data ownership	Governance process that details legal ownership of enterprise-wide data and outlines who has ability to create, edit, modify, share and restrict access to the data	13.2
machine learning	Study of algorithms and statistical models that perform a specific task without using explicit instructions, relying on patterns and inference	13.3
pattern recognition	Automated recognition of patterns and regularities in data	13.3
data mining	Process of collecting, cleaning, processing, analyzing, and discovering patterns and useful insights from large data sets	13.3
principal component analysis	Dimension reduction technique that uses orthogonal transformations to convert a set of possibly correlated variables into a set of linearly uncorrelated variables	13.3

(continued)

Term	Definition	Section
cluster analysis	Unsupervised learning method that aims to splot data into homogenous groups using a similarity measure	13.3
k-means algorithm	Type of clustering that aims to partition data into k mutually exclusive clusters by assigning observations to the cluster with the nearest centroid	13.3
linear regression	Supervised model that uses a linear function to approximate the relationship between the target and explanatory variables	13.3
generalized linear model	Supervised model that generalizes linear regression by allowing the linear component to be related to the response variable via a link function and by allowing the variance of each measurement to be a function of its predicted value	13.3
systematic component	The linear combination of explanatory variables component in a glm	13.3
link function	Function that relates between the linear predictor component to the mean of the target variable	13.3
decision trees	Modeling technique that uses a tree-like model of decisions to divide the sample space into non-overlapping regions to make predictions	13.3
categorical variable	A variable whose values are qualitative groups and can have no natural ordering (nominal) or an ordering (ordinal)	14.1
variables	A variable is any characteristics, number, or quantity that can be measured or counted.	14.1
interval variable	An ordinal variable with the additional property that the magnitudes of the differences between two values are meaningful	14.1
spatial data	Data and information having an implicit or explicit association with a location relative to the earth	14.1
high dimensional	Data set is high dimensional when it has many variables. In many applications, the number of variables may be larger than the sample size.	14.1
qualitative	This is a type of variable in which the measurement denotes membership in a set of groups, or categories	14.1
nominal variable	This is a type of qualitative/ categorical variable which has two or more categories without having any kind of natural order.	14.1

(continued)

Term	Definition	Section
ordinal variable	This is a type of qualitative/ categorical variable which has two or more ordered categories.	14.1
binary variable	Is a special type of categorical variable where there are only two categories.	14.1
quantitative variable	A quantitative variable is a type of variable in which numerical level is a realization from some scale so that the distance between any two levels of the scale takes on meaning.	14.1
continuous variable	A continuous variable is a quantitative variable that can take on any value within a finite interval.	14.1
policyholder	Person in actual possession of insurance policy; policy owner.	14.1
discrete variable	A discrete variable is quantitative variable that takes on only a finite number of values in any finite interval.	14.1
count variable	A count variable is a discrete variable with values on nonnegative integers.	14.1
circular data	In a circular data, all values around the circle are equally likely. Example, imagine an analog picture of a clock.	14.1
insurers	An insurance company authorized to write insurance under the laws of any state.	14.1
multivariate	Multivariate variable involves taking many measurements on a single entity.	14.1
workers compensation	Insurance that covers an employer's liability for injuries, disability or death to persons in their employment, without regard to fault, as prescribed by state or federal workers' compensation laws and other statutes.	14.1
univariate	Univariate analysis is the simplest form of analyzing data. "Uni" means "one", so in other words your data has only one variable.	14.1
missing data	Missing data occur when no data value is stored for a variable in an observation. Missing data can occur because of nonresponse: no information is provided for one or more items or for a whole unit or subject.	14.1
censored	Censored data have unknown values beyond a bound on either end of the number line or both. Here, the data is observed but the values (measurements) are not known completely.	14.1

(continued)

Term	Definition	Section
truncated	Truncation occurs when values beyond a boundary are either excluded when gathered or excluded when analyzed. An object can be detected only if its value is greater than some number.	14.1
stochastic process	Stochastic process is defined as a collection of random variables that is indexed by some mathematical set, meaning that each random variable of the stochastic process is uniquely associated with an element in the set.	14.1
deductibles	A deductible is a parameter specified in the contract. Typically, losses below the deductible are paid by the policyholder whereas losses in excess of the deductible are the insurer's responsibility (subject to policy limits and coninsurance).	14.1
rank based measures	Statistical dependence between the rankings of two variables	14.2
odds ratio	A statistic quantifying the strength of the association between two events, a and b, which is defined as the ratio of the odds of a in the presence of b and the odds of a in the absence of b	14.2
likelihood ratio test	A statistical test of the goodness-of-fit between two models	14.2
pearson correlation	A measure of the linear correlation between two variables	14.2
product-moment (pearson) correlation	Pearson correlation, a measure of the linear correlation between two variables	14.2
kendall tau	A statistic used to measure the ordinal association between two measured quantities	14.2
concordant	An observation pair (x,y) is said to be concordant if the observation with a larger value of x has also the larger value of y	14.2
discordant	An observation pair (x,y) is said to be discordant if the observation with a larger value of x has the smaller value of y	14.2
pearson chi-square statistic	A statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance	14.2

(continued)

Term	Definition	Section
tetrachoric correlation	A technique for estimating the correlation between two theorised normally distributed continuous latent variables, from two observed binary variables	14.2
polychoric correlation	A technique for estimating the correlation between two theorised normally distributed continuous latent variables, from two observed ordinal variables	14.2
polyserial correlation	The correlation between two continuous variables with a bivariate normal distribution, where one variable is observed directly, and the other is unobserved	14.2
biserial correlation	A correlation coefficient used when one variable is dichotomous	14.2
normal score	Transformed data which closely resemble a standard normal distribution	14.2
copula	A multivariate distribution function with uniform marginals	14.3
spearmans rho	A nonparametric measure of rank correlation	14.3
marginal distributions	The probability distribution of the variables contained in the subset of a collection of random variables	14.4
fat-tailed	A fat-tailed distribution is a probability distribution that exhibits a large skewness or kurtosis, relative to that of either a normal distribution or an exponential distribution	14.4
probability integral transformation	Any continuous variable can be mapped to a uniform random variable via its distribution function	14.4
elliptical copulas	The copulas of elliptical distributions	14.5
correlation matrix	A table showing correlation coefficients between variables	14.5
elliptical distributions	Any member of a broad family of probability distributions that generalize the multivariate normal distribution	14.5
tail dependency	A measure of their comovements in the tails of the distributions	14.5
frechet-hoeffding bounds	Bounds of multivariate distribution functions	14.5
blomqvists beta	A dependence measure based on the center of the distribution	14.7

(continued)

Term	Definition	Section
reinsurance	Insurance purchased by an insurer	1.1, 10.4
deductible	A deductible is a parameter specified in the contract. typically, losses below the deductible are paid by the policyholder whereas losses in excess of the deductible are the insurer's responsibility (subject to policy limits and coninsurance).	1.2, 5.3
coinsurance	Coinsurance is an arrangement whereby the insured and insurer share the covered losses. typically, a coinsurance parameter specified means that both parties receive a proportional share, e.g., 50%, of the loss.	1.2, 5.5
pure premium	Pure premium is the total severity divided by the number of claims. it does not include insurance company expenses, premium taxes, contingencies, nor an allowance for profits. also called loss costs. some definitions include allocated loss adjustment expenses (alae).	1.3, 7.1, 7.2
standard deviation	The square-root of variance	2.1, 3.1
variance	Second central moment of a random variable x, measuring the expected squared deviation of between the variable and its mean	2.1, 3.1
aggregate claims	The sum of all claims observed in a period of time	2.1, 5.1, 14.1
median	50th percentile of a definition, or middle value where half of the distribution lies below	3.1, 4.1
lorenz curve	A graph of the proportion of a population on the horizontal axis and a distribution function of interest on the vertical axis.	4.1, 7.6
law of total variance	A decomposition of the variance of a random variable into conditional components. specifically, for random variables x and y on the same probability space, var(x) = e[var(y|x)] + var[e(x|y)].	5.3, 9.4
tail value-at-risk	The expected value of a risk given that the risk exceeds a value-at-risk	6.2, 10.3
expected shortfall	The average value at risk	6.2, 10.3

(continued)

Term	Definition	Section
coefficient of variation	Standard deviation divided by the mean of a distribution, to measure variability in terms of units of the mean	6.3, 9.2
loss ratio	The sum of losses divided by the premium.	7.1, 7.2
homogeneous risks	Risks that have the same distribution, that is, the distributions are identical.	7.1, 7.2
heterogeneous	Heterogeneous risks have different distributions. often, we can attribute differences to varying exposures or risk factors.	7.1, 7.4
exposure	A type of rating variable that is so important that premiums and losses are often quoted on a "per exposure" basis. that is, premiums and losses are commonly standardized by exposure variables.	7.2, 7.4
loss	The amount of damages sustained by an individual or corporation, typically as the result of an insurable event.	7.5, 14.1
iid	Independent and identically distributed	
pdf	Probability density function	
aic	Akaike's information criterion	
bic	Bayesian information criterion	
pmf	Probability mass function	
mcmc	Markov Chain Monte Carlo	
cdf	Cumulative distribution function	
df	Degrees of freedom	
glm	Generalized linear model	
mle	Maximum likelihood estimate	
ols	Ordinary least squares	
pf	Probability function	
rv	Random variable	
reporting delay	The time that elapses between the occurrence of the insured event and the reporting of this event to the insurance company.	11.1
settlement delay	The time between reporting and settlement of a claim.	11.1
rbns	Reported, But is Not fully Settled	11.1
ibnr	Incurred in the past But is Not yet Reported. For such a claim the insured event took place, but the insurance company is not yet aware of the associated claim.	11.1
granular		11.1

(continued)

Term	Definition	Section
case estimates	The claims handlers expert estimate of the outstanding amount on a claim.	11.1
.csv	Comma separated value file	11.2
.txt	Text file	11.2
run-off triangle	Triangular display of loss reserve data. Accident or occurrence periods on one axis (often vertical) with development periods on the other (often horizontal). Also known as a development triangle.	11.2
development triangle	Triangular display of loss reserve data. Accident or occurrence periods on one axis (often vertical) with development periods on the other (often horizontal). Also known as a run-off triangle.	11.2
msep	Mean Squared Error of Prediction	
chain-ladder method	An algorithm for predicting incomplete losses to their ultimate cumulative value. The name refers to the chaining of a sequence of (year-to-year development) factors into a ladder of factors.	11.3
wls	weighted least squares	11.3
glm	Generalized linear model	
frequentist	Type of statistical inference based in frequentist probability, which treats probability in equivalent terms to frequency and draws conclusions from sample-data by means of emphasizing the frequency or proportion of findings in the data.	9
posterior distribution	The posterior distribution is the updated probability distribution of a parameter after incorporating prior information and observed data through Bayesian inference.	9.1
bayes' rule	A probability law that expresses conditional probability of the event a given the event b in terms of the conditional probability of the event b given the event a and the unconditional probability of a	9.1
informative	An informative prior, in statistics, is a prior probability distribution that is chosen deliberately to incorporate specific information or beliefs about a parameter before observing new data.	9.2

(continued)

Term	Definition	Section
weakly informative	A weakly informative prior is a prior probability distribution that introduces some general constraints or vague beliefs about a parameter, without heavily influencing the final inference.	9.2
noninformative	A noninformative prior is a prior probability distribution that intentionally avoids incorporating specific information or strong beliefs about a parameter.	9.2
improper	An improper prior is a prior probability distribution that does not integrate to a finite value over the entire parameter space.	9.2
conjugate distributions	Conjugate distributions are specific pairs of prior and likelihood functions that result in a posterior distribution within the same family of probability distributions as the prior.	9.3
hyperparameters	Hyperparameters are parameters that define the distribution of a prior distribution	9.3
gibbs sampler	The Gibbs sampler is an iterative algorithm in statistics used for simulating samples from complex probability distributions. It's particularly useful in Bayesian analysis for drawing samples from multivariate distributions by updating one variable at a time while keeping others fixed.	9.4
metropolis–hastings algorithm	The Metropolis–Hastings algorithm is a method to generate samples from complex distributions by proposing new samples and deciding whether to accept them, making it valuable for Bayesian analysis and complex modeling.	9.4
precision	Precision is the inverse of variance and is often used to quantify the amount of uncertainty or variability in a prior or posterior distribution.	9.3